Transforming Text Styles: Leveraging Fine-Tuned T5 Models for Bidirectional Style Transfer

Tilak Matagunde
PES University
tilakmatagunde391@gmail.com

Pushparaj Shetty K S
PES University
kspushparajshetty@gmail.com

Praneeth Cheepurupalli
PES University
chpraneeth2003@gmail.com

Dr. Preethi P
PES University
preethip@pes.edu

Dr. Mamatha H R
PES University
mamathahr@pes.edu

***Abstract*—Style transfer in natural language processing (NLP) has emerged as a critical area of research, facilitating the transformation of text between various styles such as formal to informal, active to passive, and positive to negative. This paper explores the methodologies and challenges associated with these transformations and presents novel approaches to address them. We investigate the significance of style transfer in enhancing text generation tasks and discuss its applications across different domains. Through experimental evaluations, we demonstrate the effectiveness of our proposed techniques in achieving accurate and coherent style transformations.**

***Index Terms*—Style transfer, Natural language processing (NLP), Formal to informal conversion, Active to passive transformation, Positive to negative sentiment alteration, Text generation, Applications, Methodologies, Challenges, Experimental evaluations**

I. Introduction

In the realm of natural language processing (NLP), style transfer stands as a transformative technique, allowing for the alteration of text styles while retaining the core semantic meaning. This entails the conversion of text from one stylistic form to another, encompassing tasks such as shifting from formal language to informal expressions, from active voice constructions to passive constructions, and from positive sentiment to negative sentiment and their inverse. The implications of such transformations are profound, extending across various NLP applications including but not limited to text generation, sentiment analysis, and paraphrasing. By facilitating the adaptation of text to diverse contexts and audiences, style transfer serves as a catalyst for enhanced communication and personalized content delivery. Nonetheless, the pursuit of accurate and fluent style transformations is rife with challenges. These challenges include the necessity of preserving content fidelity amidst stylistic alterations, grappling with syntactic disparities between different styles, and ensuring coherence and naturalness in the resulting text.

II. Dataset

For our experiments on bidirectional text style transfer, we utilize diverse datasets that cater to different transformation tasks. Among these datasets, the Yelp dataset plays a crucial role in facilitating bidirectional style transfer between positive and negative sentiments.

A. Yelp Dataset

he Yelp dataset is a widely used resource in natural language processing, particularly for sentiment analysis tasks. It consists of user reviews collected from the Yelp platform, annotated with sentiment labels indicating whether the review expresses a positive or negative sentiment. The dataset encompasses a diverse range of domains, including restaurants, hotels, and various services, providing a rich and varied corpus for sentiment analysis and text style transfer tasks.

B. GYAFC Datadet

The GYAFC dataset, also known as the Formal-Informal Corpus, is a valuable resource for studying bidirectional style transfer between formal and informal language styles. It comprises parallel data pairs consisting of sentences in both formal and informal styles, sourced from various online forums, social media platforms, and written sources. Each parallel pair consists of a sentence in formal language and its corresponding informal variant, enabling the training and evaluation of models for bidirectional style transfer tasks.

III. Related Work

A rich tapestry of research endeavors has been woven around the methodologies and techniques employed in style transfer within the domain of NLP. Early forays into this domain often relied on rule-based systems, wherein predefined patterns or templates governed the transformation process. However, such methods were often rigid and struggled to generalize across diverse styles and domains. The advent of neural network-based models marked a paradigm shift in this landscape, ushering in unprecedented advancements in unsupervised and semi-supervised style transfer. Pioneering works by Sennrich et al. (2016) introduced sequence-to-sequence architectures, showcasing their efficacy in seamlessly transferring text between different languages and stylistic realms. Li et al. (2018) further pushed the boundaries with a reinforcement learning-based approach, affording fine-grained control over stylistic attributes during text generation. Moreover, hybrid systems amalgamating rule-based and neural network-based methodologies have emerged as potent contenders, promising more robust and adaptable style transformations. Despite

these strides, the challenge of preserving content relevance, grappling with syntactic nuances, and ensuring fluency across disparate styles and domains remains a formidable hurdle.

IV. Problem Statement Formulation

This research endeavors to confront the multifaceted challenges inherent in style transfer within the purview of NLP, with a concerted focus on four pivotal transformations: the conversion of formal language to informal vernacular, the transition from active voice constructions to passive voice constructions, the modulation of sentiment polarity from positive to negative, and their inverse. Our overarching objective is to devise methodologies capable of effecting accurate and fluent transformations while steadfastly preserving the semantic integrity of the text and ensuring coherence in the resultant output. Central to this pursuit are the thorny issues surrounding syntactic variations between divergent styles, the imperative of maintaining content fidelity throughout stylistic metamorphoses, and the judicious control over the degree of style manipulation to yield desired outcomes. Additionally, we aim to probe the reverberations of style transfer on downstream NLP tasks, such as text generation and sentiment analysis, and subject our proposed methodologies to rigorous scrutiny through comprehensive experimental evaluations.

V. proposed metholodogy

In this section, we outline the methodology employed for text sentiment transfer, focusing on the fine-tuning of T5-Small and T5-Base models. These models serve as the backbone for our bidirectional sentiment alteration tasks, enabling the transformation of text between positive and negative sentiment polarities.

A. T5 Model Architecture

The Text-To-Text Transfer Transformer (T5) model, introduced by Raffel et al. (2019), is a versatile architecture that excels in various natural language processing tasks, including text generation, translation, and summarization. T5 adopts a unified text-to-text framework, wherein both input and output are represented as text strings, allowing for seamless integration of different tasks into a single model.

B. T5-Small

The T5-Small variant is a scaled-down version of the T5 model, characterized by a smaller number of layers and parameters compared to its larger counterparts. Despite its reduced size, T5-Small retains the core architecture and functionality of the original T5 model, making it suitable for tasks requiring moderate computational resources while maintaining competitive performance.

C. T5-Base

The T5-Base variant represents the baseline configuration of the T5 model, featuring a moderate number of layers and parameters optimized for a balance between computational efficiency and task performance. T5-Base serves as a versatile choice for a wide range of natural language processing tasks, offering a favorable trade-off between model size and performance.

D. Fine-Tuning for Sentiment Transfer

To enable bidirectional sentiment alteration, we fine-tune both the T5-Small and T5-Base models on datasets containing pairs of sentences with opposing sentiment polarities. The fine-tuning process involves updating the model parameters to adapt to the specific task of sentiment transfer while leveraging the pre-trained representations learned from large-scale text corpora. During fine-tuning, we initialize the T5-Small and T5-Base models with pre-trained weights obtained from the original T5 model. We then train the models on the sentiment transfer dataset using supervised learning techniques, where each input-output pair consists of a sentence with its corresponding sentiment-altered counterpart. The objective is to minimize the loss function, typically cross-entropy loss, between the generated and target sentences.

E. Formal-to-Informal and Informal-to-Formal Style Transfer

The GYAFC dataset facilitates bidirectional style transfer between formal and informal language styles, allowing models to learn to convert text seamlessly between these two variants. By leveraging this dataset, we can train models capable of transforming formal language into informal vernacular and vice versa. This enables enhanced communication flexibility and adaptation of text to diverse contexts and audiences, catering to a wide range of linguistic preferences and scenarios.

The utilization of the GYAFC dataset complements the Yelp dataset by providing additional linguistic diversity and context for training and evaluating models for formal-to-informal and informal-to-formal style transfer tasks. By harnessing the rich parallel data pairs available in the GYAFC dataset, we can develop robust models capable of bidirectional style transfer across different language styles, further enriching the capabilities of text generation and adaptation in natural language processing applications.

VI. Experiments and Results

In this section, we describe the experiments conducted to compare the performance of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for the classification of AI-generated images. We present the experimental setup, including dataset preparation, model architectures, training procedures, and evaluation metrics. Subsequently, we provide a detailed analysis of the obtained results.

A. Experimental Setup

B. Dataset

Our project utilizes diverse datasets to train and evaluate models for bidirectional text style transfer across multiple dimensions. Each dataset serves as a valuable resource, providing linguistic diversity and context for robust model training and evaluation.

1) Formal-to-Informal and Informal-to-Formal Style Transfer: For bidirectional style transfer between formal and informal styles, we employ the Gyafc dataset (Formal-Informal Corpus). This dataset contains parallel data pairs consisting of sentences in both formal and informal styles, sourced from various online forums and social media platforms. Leveraging this dataset allows us to train models capable of transforming text seamlessly between formal and informal variants, enhancing communication flexibility.

2) Active-to-Passive and Passive-to-Active Style Transfer: To address bidirectional style transfer between active and passive voice constructions, we utilize the "Active/passive sentence pairs and their embeddings" dataset. This dataset comprises pairs of sentences, with one sentence in the active voice and the other in the passive voice, along with their corresponding embeddings. By leveraging this dataset, we train models capable of bidirectional transformation between active and passive voice constructions, catering to diverse linguistic contexts and preferences.

3) Positive-to-Negative and Negative-to-Positive Style Transfer: For bidirectional style transfer between positive and negative sentiments, we leverage parallel data extracted from the Yelp dataset. This dataset offers a rich collection of user reviews annotated with sentiment labels, allowing us to construct pairs of reviews with opposing sentiment polarities. By training models on this dataset, we enable bidirectional sentiment alteration in natural language text, facilitating nuanced expression and communication across different domains.

4) Model Architectures:

- **Formal to Informal:** T5-Small model fine-tuned on a dataset for converting formal language to informal language.
- **Informal to Formal:** T5-Small model fine-tuned on a dataset for converting informal language to formal language.
- **Active to Passive Voice:** T5-Small model fine-tuned on a dataset for converting active voice sentences to passive voice sentences.
- **Passive to Active Voice:** T5-Small model fine-tuned on a dataset for converting passive voice sentences to active voice sentences.
- **Positive to Negative:** T5-Base model fine-tuned on a dataset for converting positive statements to negative statements.
- **Negative to Positive:** T5-Base model fine-tuned on a dataset for converting negative statements to positive statements.

5) Training Procedure: For all models, the training procedure involved fine-tuning the pre-trained T5 model on the respective datasets using the following training parameters:

- Learning Rate: 2×10^{-5}
- Batch Size: 16
- Weight Decay: 0.01
- Number of Epochs: Varied (e.g., 3-5 epochs)

6) Evaluation Metrics: The performance of each model was evaluated using the following metrics:

- Loss: Calculated using the cross-entropy loss function.
- Rouge1, Rouge2, Rougel, and Rougelsum: Evaluation metrics for text generation tasks, measuring the overlap between generated and reference text.
- Gen Len: Mean generated length of the output text.

C. Results

The results of the experiments are summarized below:

- **Formal to Informal:** Loss: 0.5109, Rouge1: 84.3715, Rouge2: 72.1078, Rougel: 84.2884, Rougelsum: 84.2975, Gen Len: 14.2801
- **Informal to Formal:** Loss: 0.7186, Rouge1: 80.8589, Rouge2: 67.0673, Rougel: 80.7564, Rougelsum: 80.7571, Gen Len: 14.9451
- **Active to Passive Voice:** Loss: 0.1036, Rouge1: 96.7196, Rouge2: 94.1746, Rougel: 95.2986, Rougelsum: 95.3129, Gen Len: 16.5466
- **Passive to Active Voice:** Loss: 0.1232, Rouge1: 97.9812, Rouge2: 94.9546, Rougel: 95.8714, Rougelsum: 95.8973, Gen Len: 14.9464
- **Positive to Negative:** Loss: 0.1630, Rouge1: 66.0728, Rouge2: 48.2651, Rougel: 65.0881, Rougelsum: 64.9953, Gen Len: 14.0178
- **Negative to Positive:** Loss: 0.1739, Rouge1: 62.4361, Rouge2: 46.2313, Rougel: 60.8217, Rougelsum: 60.8436, Gen Len: 14.2644

D. Conclusion

In this study, we explored the effectiveness of fine-tuned T5-Small and T5-Base models for various text style transfer tasks, including formal-to-informal conversion, informal-to-formal conversion, active-to-passive voice transformation, passive-to-active voice transformation, positive-to-negative sentiment alteration, and negative-to-positive sentiment alteration. Our experiments revealed promising results across all evaluated tasks, demonstrating the capability of the fine-tuned T5 models to effectively perform bidirectional style transfer while maintaining content fidelity and linguistic coherence.

For formal-to-informal and informal-to-formal style transfer tasks, the fine-tuned T5-Small model achieved competitive performance, as evidenced by high Rouge scores and low loss values. Similarly, the T5-Small model exhibited excellent performance in transforming between active and passive voice constructions, achieving high Rouge scores and minimal loss. Moreover, the T5-Base model demonstrated proficiency in sentiment alteration tasks, successfully converting between positive and negative sentiment polarities with satisfactory Rouge scores and loss values.

The consistent performance of the fine-tuned T5 models across diverse style transfer tasks underscores their versatility and effectiveness in natural language processing applications.

By leveraging pre-trained language representations and fine-tuning on task-specific datasets, we can harness the power of transfer learning to achieve accurate and fluent style transformations.

Moving forward, further research could explore enhancements to the fine-tuning process, such as experimenting with different hyperparameters, model architectures, and training strategies. Additionally, investigating the generalization capabilities of the fine-tuned models across various domains and linguistic contexts would provide valuable insights into their robustness and adaptability in real-world scenarios.

Overall, our study contributes to the advancement of text style transfer techniques and highlights the potential of fine-tuned T5 models as powerful tools for linguistic adaptation and communication in natural language processing applications.

REFERENCES

[1] Manakul, P., Liusie, A.,Gales, M.J. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. ArXiv, abs/2303.08896.

[2] Friel, R.,Sanyal, A. (2023). Chainpoll: A high efficacy method for LLM hallucination detection. ArXiv. /abs/2310.18344

[3] Yang, S., Sun, R., Wan, X. (2023). A New Benchmark and Reverse Validation Method for Passage-level Hallucination Detection. ArXiv. /abs/2310.06498

[4] Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., Dudek, G. (2024). Hallucination Detection and Hallucination Mitigation: An Investigation. ArXiv. /abs/2401.08358

[5] Yao, J., Ning, K., Liu, Z., Ning, M., Yuan, L. (2023). LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. ArXiv. /abs/2310.01469

[6] Towhidul, S Tonmoy, S M Zaman, S M Mehedi Jain, Vinija Rani, Anku Rawte, Vipula Chadha, Aman Das, Amitava. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. 10.13140/RG.2.2.11724.39045.

[7] Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., Ye, J. (2024). INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. ArXiv, abs/2402.03744.

[8] Wei, C., Chen, Z., Fang, S., He, J., Gao, M. (2024). OPDAI at SemEval-2024 Task 6: Small LLMs can Accelerate Hallucination Detection with Weakly Supervised Data. ArXiv, abs/2402.12913.